Handling Missing Annotations for Semantic Segmentation with Deep ConvNets

Olivier Petit[1,2(✉)], Nicolas Thome[1], Arnaud Charnoz[2], Alexandre Hostettler[3], and Luc Soler[2,3]

[1] CEDRIC - Conservatoire National des Arts et Metiers, Paris, France
[2] Visible Patient SAS, Strasbourg, France
olivier.petit@visiblepatient.com
[3] IRCAD, Strasbourg, France

Abstract. Annotation of medical images for semantic segmentation is a very time consuming and difficult task. Moreover, clinical experts often focus on specific anatomical structures and thus, produce partially annotated images. In this paper, we introduce SMILE, a new deep convolutional neural network which addresses the issue of learning with incomplete ground truth. SMILE aims to identify ambiguous labels in order to ignore them during training, and don't propagate incorrect or noisy information. A second contribution is SMILEr which uses SMILE as initialization for automatically relabeling missing annotations, using a curriculum strategy. Experiments on 3 organ classes (liver, stomach, pancreas) show the relevance of the proposed approach for semantic segmentation: with 70% of missing annotations, SMILEr performs similarly as a baseline trained with complete ground truth annotations.

Keywords: medical images · deep learning · convolutional neural networks · incomplete ground truth annotation · noisy labels · missing labels.

1 Introduction

Fully automatic semantic segmentation of medical images is a major challenge. Over the last few years, Deep Learning and Convolutional Neural Networks (ConvNets) have reached outstanding performances on various visual recognition tasks [9]. Regarding semantic segmentation on natural images, state-of-the-art performances are currently obtained with Fully Convolutional Neural Networks (FCNs) [1, 3]. Consequently, several attempts have been made to apply those methods on medical images [15, 11, 16]. In challenges like Liver Tumor Segmentation Challenge (LiTS), leading methods are based on FCNs [5, 10].

However, training deep ConvNets requires large amount of data with clean annotations. The annotation process is an extremely time consuming task for semantic segmentation, which requires pixel-level labeling. This challenge is amplified in the medical field, where highly qualified professionals are needed. In this paper, we focus on abdomen 3D CT-scans from an internal dataset with more than 1000 patients, each volume containing about a hundred of 512×512

images. The segmentation masks have been realized by clinical experts but they have focused on specific organs or anatomical structures, *e.g.* liver pathologies. As a consequence, the collected labels intrinsically contain missing annotations, as illustrated in Figure 1.

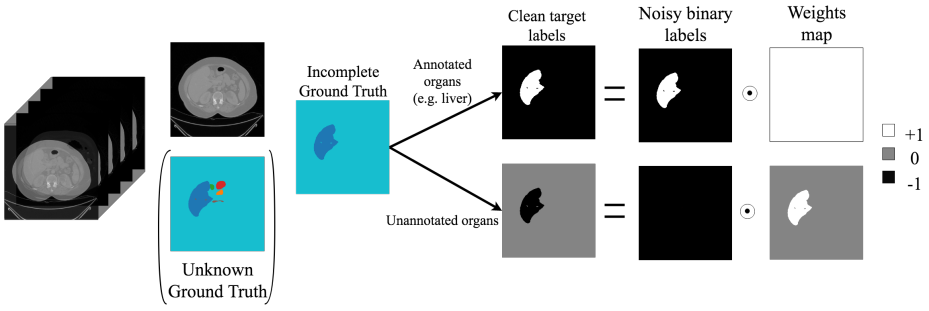


Fig. 1: Our 3D CT-scan dataset is labeled by clinical experts who focused on certain organ pathologies, *e.g.* liver. The ground truth annotations are therefore incomplete. We define ambiguity maps to train binary class predictors, which ignore incorrect background labels.

Several learning methodologies can be used to address the aforementioned missing annotations issue. Weakly Supervised Learning (WSL) can be used to leverage coarse annotations, *e.g.* global image or volume labels. WSL is generally closely connected to Multiple Instance Learning [4], and has been used for WSL segmentation of natural images [14, 13] and medical data [7]. However, performing pixel-wise prediction from global labels is known to be a challenging task, making WSL approaches generally substantially inferior to their fully supervised counterparts. Since missing annotations are incorporated to background pixel classes, another option to address this problem is to design models able to incorporate noisy labels, which have been recently applied for semantic segmentation [12, 8]. Although interesting, most of these methods rely on the assumption that the ratio of noisy labels remains relatively low, whereas more than 50% of the organs are commonly missing in our context.

In this paper, we introduce SMILE, a new method for Semantic segmentation with MIssing Labels and ConvNEts. Firstly, we design a learning scheme which converts the segmentation of K organ classes into K binary problems, and we define ambiguity maps which allow to train the model with 100% of clean labels (see Figure 1), while retaining a largely sufficient number of negative samples. The model trained at this first stage is then used for automatically predicting labels for missing organs, using a Curriculum strategy [2] (SMILEr). We perform extensive experiments in an sub-set of our dataset for the segmentation of three organ classes: liver, pancreas and stomach. We show that our approach significantly outperform a strong FCN baseline based on Deeplab [3], especially when the number of missing organs is large. The final model (SMILEr) trained

with only 30% of present organs performs similarly to a baseline trained with complete ground truth annotations.

2 SMILE Model

The SMILE model is dedicated to semantic segmentation with missing labels using ConvNets. The missing organ annotations are labeled as "background", as shown in Figure 1.

SMILE is based on the strong DeepLab baseline [3], which shows impressive results for natural and medical images [5]. The DeepLab backbone architecture is a Fully Convolutional Networks (FCN), as shown in Figure 2, *e.g.* Res-Net [6]. In DeepLab, 1x1 convolutions and soft-max are applied to classify each pixel into K (+1, *i.e.* background) classes.

2.1 Handling missing annotations

In our context, the main limitation of DeepLab is that background labels sometimes correspond to missing organs. Therefore, back-propagating these background labels may damage training performances by conflicting with pixels where the organ is properly annotated.

SMILE architecture To address this problem, we choose to start from the $(K+1)$ multi-class classification formulation, and to classify each organ independently using K binary classifiers. The SMILE architecture is shown in Figure 2. We use 1×1 convolutions, as in DeepLab, but we apply a sigmoid activation function to predict the presence / absence of an organ at each pixel.

SMILE training During training, the K binary models generate K losses at each pixel by computing the binary cross entropy: $L_k(\hat{y_k}, y_k^*) = -(y_k^* \; log(\hat{y_k}) + (1-y_k^*) \; log(1-\hat{y_k}))$. The final loss aggregates these K losses through summation:

$$L(\hat{y}, y^*) = \sum_{k=1}^{K} w_k \; L_k(\hat{y_k}, y_k^*) \tag{1}$$

where $w_k \in \{0; 1\}$ is a binary weight map which select or ignore pixels for class k.

The w_k weights are the core of the SMILE model, which are used to ignore ambiguous annotations during training. We illustrate the rationale of our approach in Figure 2. We consider a volume where only one organ is annotated. In the baseline DeepLab model, pixels for the other organs in each slice are incorrectly labeled as background, and back-propagated consequently. Contrarily, with SMILE, we only back-propagate labels which are certain. In this example, we can back-propagating positive / negative labels for the annotated organ at every pixels p: we thus have $w_a = 1 \forall p$. On the other hand, for unannotated

organs, we only use pixels which are certainly not belonging to the given class for training the binary classifier: $w_u = 1$ for all pixels of the annotated organs. Other pixels are ignored during training, *i.e.* $w_u = 0$.

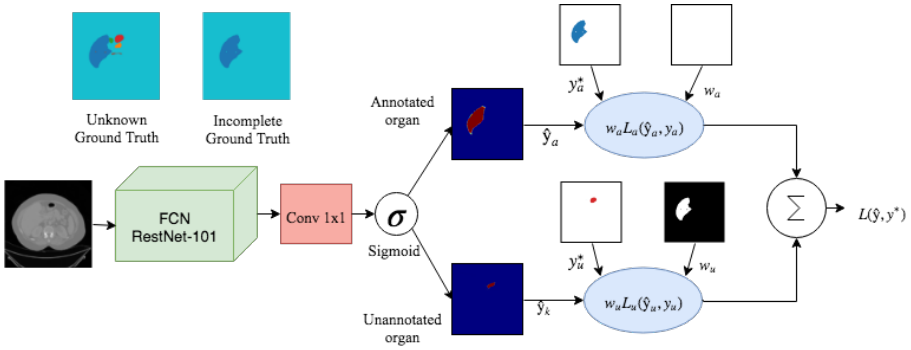


Fig. 2: SMILE architecture and training. The presence of an organ at each pixel is determined by using K independent binary classifiers. During training, a weight w_k for each class enables to ignore ambiguous pixels.

The idea behind SMILE is to only use true positive and true negative labels during training. To formalize this, we consider a given organ class k with its associated binary classification problem. We denote as β_k the ratio of pixels for the organ in all volumes of image slices, and α the ratio of missing labels for this organ in the dataset. Table 1 shows the confusion matrix for the labels used by SMILE and the DeepLab baseline. We can see that they both use the same amount of true positives: $TP = (1 - \alpha) \cdot \beta_k$. For negative examples, however, the baseline uses $FN = \alpha \cdot \beta_k$ false negatives, *i.e.* the amount of unannotated pixels belonging to the organ. The ratio $\frac{TP}{FN} = \frac{1-\alpha}{\alpha}$ gives a good indication on the influence of the wrong information: with $\alpha > 0.5$, $\frac{TP}{FN} < 1$, which means that the model incorporates more wrong labels than correct ones, dramatically deteriorating its performances.

On the other hand, the baseline learns with more true negatives $(1 - \beta_k)$ than SMILE $(1 - \alpha)(1 - \beta_k) + \epsilon$, where $\epsilon = \sum_{k' \neq k} \beta_{k'}$ corresponds to the other organ labels (see Figure 2). However, we take advantage on the class unbalance: generally $\beta << 1$, *e.g.* $\beta = 0.05$, since the organs represent a small proportion of the total volume. As a consequence, even if we remove some background examples, we still have largely enough information to learn it properly.

2.2 Incremental self-supervision and relabeling

The number of true positives (TP) is linearly decreasing with respect to the ratio of missing organ annotation α (Table 1). SMILE can thus be improved by recovering TP in unannotated training images. We propose a self-supervised

Table 1: Training label analysis. GT: Ground Truth

(a) Baseline FCN

GT \ Used	Pos	Neg
Pos	$(1-\alpha)\cdot\beta_k$	$\alpha\cdot\beta_k$
Neg	0	$1-\beta_k$

(b) SMILE

GT \ Used	Pos	Neg
Pos	$(1-\alpha)\cdot\beta_k$	0
Neg	0	$(1-\alpha)\cdot(1-\beta_k)+\epsilon$

approach to achieve this goal, called SMILEr (SMILE with relabeling). The idea of SMILEr is to iteratively produce new positive target labels $y^*_{i,t} = 1$ in an image with missing annotations $\mathbf{x}_i$ for each class k[4], using a curriculum strategy [2].

Basically, SMILEr is initialized with SMILE, which has been trained with correct positive labels only (Table 1) that can be regarded as "easy positive samples". Let us denote as $\hat{y_i}^+$, the pixels predicted as positive by SMILE in a given unannotated image $\mathbf{x}_i$. SMILEr then add new $\oplus$ labels $y^{*,+}_{i,t}$ by selecting the top scoring pixels among $\hat{y_i}^+$. The model is then retrained with the augmented training set, and the process is iterated T times, by selecting an increasing ratio $\gamma_t = \frac{t}{T}\gamma_{max}$ of top scoring pixels among positives.

The new $\oplus$ labels $y^*_{i,t}$ incorporated at each curriculum iteration are "harder examples", since they are incrementally determined by the model trained with an increased set of auto-supervised positives.

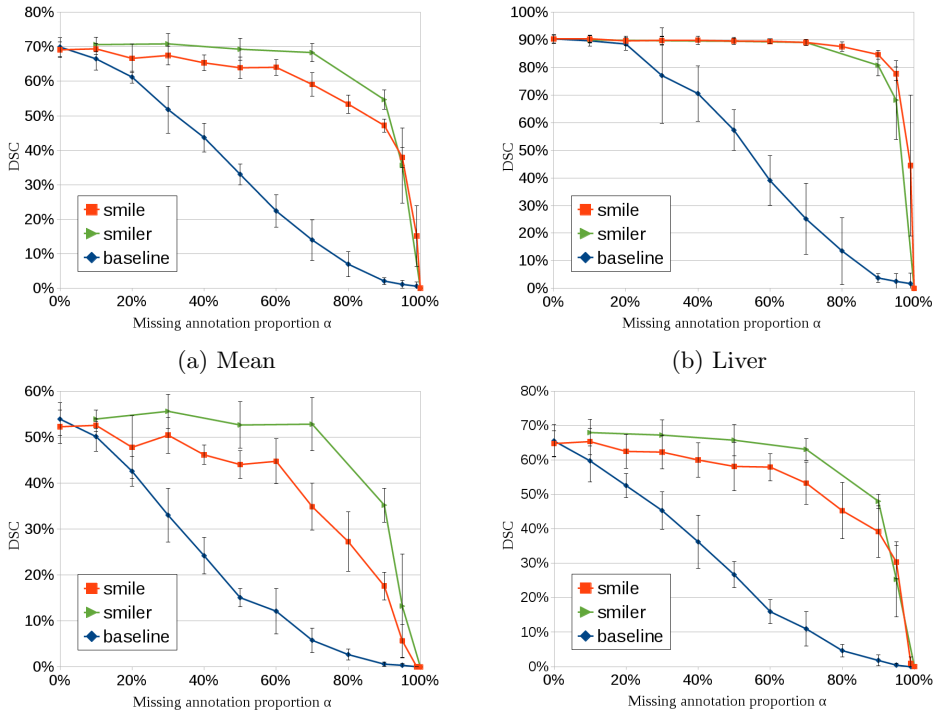
3 Experiments and Results

We perform experiments on a subset of our dataset with complete ground truth annotations for three organs: liver, pancreas and stomach, which gathers 72 3D volume CT-scans. We generate a partially annotated dataset by randomly removing α% of organs in the volumes independently.

Quantitative evaluations We compare our approach to the DeepLab baseline [3] with a varying ratio of missing annotations α. We randomly split training (80%) and testing (20%) data K times, and report averages and standard deviations of Dice scores over the K runs. For SMILEr, we fix $T = 2$ and $\gamma_{max} = 0.66$.

Figure 3 shows the results for the baseline, SMILE and SMILEr, for each organ and on average. As expected, the maximum scores are reached when 100% of the annotations are kept, *i.e.* $\alpha = 0$. When α increases, the performances of the baseline dramatically drop, whereas our approach continues to perform well. For example, SMILE performs similarly as the method trained with complete annotations with $\alpha = 40$%, whereas the baseline performance is decreased by about 20 points. The gain is even more pronounced for SMILEr which results

[4] We drop the dependence of class in $y^*_{i,t}$ for clarity.



(a) Mean (b) Liver

(c) Pancreas (d) Stomach

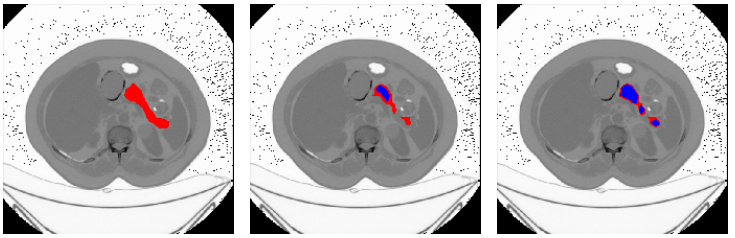
Fig. 3: Dice score versus the proportion of missing annotations α. The baseline is represented in blue, SMILE in red and SMILEr in green.

are comparable to the fully annotated method for $\alpha = 70\%$, whereas the baseline performs very poorly in this regime.

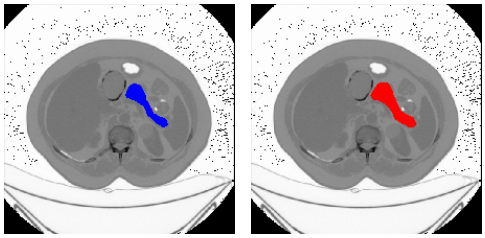
SMILEr analysis Figure 3 highlights the fact that the Dice score is better when the organ is bigger. Regarding SMILEr, we can observe that its improvement is especially pronounced for small organs, see for example the large performance boost for pancreas and stomach.

Figure 4 shows how the training evolves during the $T = 3$ curriculum iterations of SMILEr, and with $\gamma_{max} = 1$. At $t = 0$, we show the segmentation of SMILE, blue pixels indicating the new positive labels added for training for the next step. We can see how the segmentation is refined and is nearly perfect at $\gamma_2 = 0.66$ ($t = 2$). It is also interesting to see how the model tends to over predict some labels at $\gamma_3 = 1.0$.

Finally, we give in Figure 5 the final segmentation for the three organ classes in a test image, for SMILEr and the baseline, at $\alpha = 70\%$. We can notice the incapacity of the baseline, whereas SMILEr successfully segments all organs.

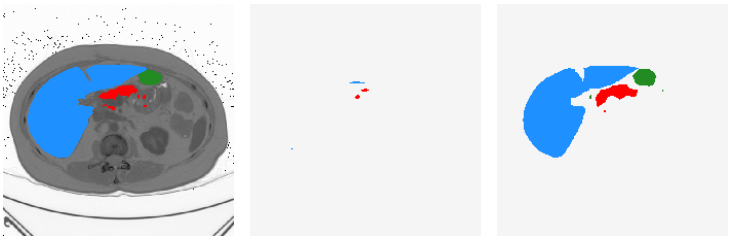


(a) Ground truth (b) $t = 0$ (c) $t = 1$



(f) $t = 2$ (g) $t = 3$

Fig. 4: SMILEr behaviour with $T = 3$ iterations, $\gamma_{max} = 1.0$ and $\alpha = 50\%$. SMILEr prediction in red, selected $\oplus$ pixels for the next iteration in blue.



(a) Ground Truth (b) Baseline $\alpha = 70\%$ (c) SMILEr $\alpha = 70\%$

Fig. 5: Segmentation results for the baseline and SMILEr, with $\alpha = 70\%$. The liver is in blue, the pancreas in red and the stomach in green.

4 Conclusions

We introduce a new model, SMILE, dedicated to semantic segmentation with incomplete ground truth. SMILE is based on the use of certain labels for training a first model, which is lately used to incrementally re-label positive pixels. Experiments show that SMILE can achieve comparable performances to a model trained with complete annotations with only 30% of labels. Future works are the application of SMILE to other organ classes, and the incorporation of uncertainty for selecting the target pixels labels in our curriculum approach.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 41–48. ICML '09 (2009)
3. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
4. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1-2), 31–71 (Jan 1997)
5. Han, X.: Automatic liver lesion segmentation using A deep convolutional neural network method. CoRR **abs/1704.07239** (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778 (2016)
7. Hwang, S., Kim, H.: Self-transfer learning for weakly supervised lesion localization. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II. pp. 239–246 (2016)
8. Kraus, O.Z., Ba, L.J., Frey, B.J.: Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics **32**(12), 52–59 (2016)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
10. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C., Heng, P.: H-denseunet: Hybrid densely connected unet for liver and liver tumor segmentation from CT volumes. CoRR **abs/1709.07330** (2017)
11. Liu, H., Feng, J., Feng, Z., Lu, J., Zhou, J.: Left atrium segmentation in ct volumes withfully convolutional networks. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 39–46. Springer International Publishing, Cham (2017)
12. Lu, Z., Fu, Z., Xiang, T., Han, P., Wang, L., Gao, X.: Learning from weak and noisy labels for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(3), 486–500 (March 2017)
13. Mordan, T., Durand, T., Thome, N., Cord, M.: WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Localization and Segmentation. In: Computer Vision and Pattern Recognition (CVPR) (2017)
14. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? Weakly-supervised learning with convolutional neural networks. In: CVPR (2015)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI (3). Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015)
16. Trullo, R., Petitjean, C., Ruan, S., Dubray, B., Nie, D., Shen, D.: Joint segmentation of multiple thoracic organs in CT images with two collaborative deep architectures. MICCAI'17 workshop Deep Learning in Medical Image Analysis (2017)